Topic Modeling on US Airline Tweets to Compare Airline Marketing Strategies

Raiha Khan
UNI: rk3148

Abstract

The airline industry has come a long way thanks to several aviation-related and technological innovation, all while catering to the needs of the modern traveler. As most large businesses that have anything to promote, airline companies market to prospective customers through Twitter, by keeping travelers aware of relevant travel information, the latest deals they may have, or a unique offering that may distinguish them from other airlines. In this paper, we use the topic modeling technique of Latent Dirichlet Allocation (LDA) to discover how airline companies use Twitter to drive consumer and traveler engagement; the analysis demonstrates airline companies' shared emphasis on promoting i) investments in charitable causes, ii) the best (or lowest) fares, or iii) the airline's growth as a company, among many other topics.

1. Introduction

Airlines are interested in how they can best support their customer base and prospective travelers by establishing a two-way communication between them on social media. Through these efforts, one scrolling on their Twitter feed can be converted into an engaged traveler and activated member of the airline's customer base. Incorporating innovative ways to help customers better understand what an airline offers and how the airline can serve them during their travel experience can help airlines build customer loyalty as well as promote their offerings via online reviews/forums or simple word-of-mouth exchanges that have the potential to take place between engaged customers and prospective travelers. For example, airlines build a near-real-time customer service pipeline to answer typical travel questions quickly [1], tailor their social media presence to attract specific customer segments (i.e. younger generations, residents of nearby countries) [2] [3], or discover new opportunities for marketing campaigns to ultimately build brand awareness of the airline in the customer [4].

Therefore, deriving the (extent to which) certain topics make up airline companies' tweets can tell us what topics certain airlines focus on more than others; from this, we can determine what is important to prospective travelers on an airline-by-airline basis. For example, one airline might especially promote their exclusive club membership benefits, while another may focus on marketing to customers that are looking for the cheapest airfare.

One way to understand where airline companies stand in their marketing strategies is by aggregating their one-way airline-to-customer communications (which, in this project, are represented as organic tweets and retweets) to perform topic modeling. Topic modeling is a text mining technique used in research to derive hidden topics and meaningful semantic structures from text across a corpus of documents.

This paper discusses related motivations and research regarding marketing strategies in the airline industry, where the research has widely focused on one-way *customer-to-airline* engagement, while the work discussed in this paper is more aligned with *airline-to-customer* engagement.

What distinguishes this project is it seeks to assess not customer engagement/experience, as in the examples referred to above, but rather *company* engagement. As such, airline companies can benefit from visual results this paper's work derives to better visualize where their social media strategy focus is currently directed, in order for them to, for example:

- **Justify the creation or adjustment of marketing strategy**: Determine where resources supporting airlines' marketing/social media strategies need to be strengthened or relocated, which can be seen as necessary to do before infusing customer feedback analyses into campaigns, which may need directional and financial backing before approval.
- **Perform competitor-topic analyses**: See where competitor airline companies lie on spectrums pertaining to similar or new/emerging topics in their tweets.

2. Related work

2.1. Analyses of tweets related to airline companies

A variety of academic research involve analyzing Twitter accounts or tweets related to airline companies.

Punel et. al. segment customers that follow or interact with an airline's Twitter account to better understand the airline's customer base [5]. Such an analysis can be layered in with this project on topic-modeling tweets from the Twitter accounts of airline companies, by aggregating customer engagement with specific topics at the segment level. Rane et. al. perform sentiment analysis on airline-specific customer feedback from tweets [6]. This can help airlines understand travelers' pain points, which they can potentially address as part of the travel experience or as making a commitment to improving their social media strategy.

Overall, these analyses can promote ideas to help airlines improve their business and marketing strategies by tailoring to their travelers' needs and attracting more customer segments in the process.

2.2. Usage of LDA for topic modeling on tweets

LDA [7] is a topic modeling method that supports analyzing a large set of documents. It takes as input a document corpus, from which a document-term matrix representation is derived for further computation. LDA then produces two different matrices: a document-topic matrix that describes each document's composition with respect to the derived topics, and a topic-word matrix that describes the likelihoods of each word being associated with each topic. In light of our context, the resulting document-topic matrix from the LDA process can help to understand to what extent a particular tweet/set of tweets is/are associated with each topic, and the resulting topic-term matrix can help to dig deeper into what terms and to what extent terms make up a topic.

Topic modeling has been used in a wide variety of tweet analysis research as a way to uncover hidden meanings beneath tweets, which are the documents in this context [8]. In particular, Kwon et. al. have explored topic modeling in relation to airline online reviews, lending more insight into customer-to-airline engagement and satisfaction [9].

3. Data

All code development for this project so far was completed in the Python programming language and utilizing the PyCharm and Jupyter Lab IDEs; all code can be found on GitHub[1].

3.1. Data downloading and acquisition

Organic tweets and retweets were collected from 85 Twitter accounts associated with airline companies headquartered worldwide using Twitter V2 API[2]. For a given Twitter handle corresponding to an airline (i.e. @united, the Twitter handle for United Airlines), the following steps were taken to download its tweets:

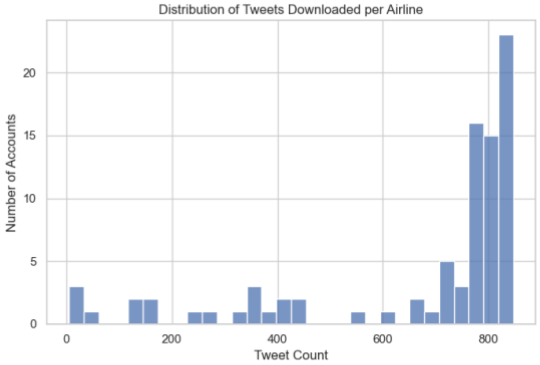


Figure 1: Distribution of tweets downloaded per airline.

1. Get the user ID of the airline company's Twitter account (unique ID associated with one Twitter account) using the user lookup endpoint (i.e. 260907612, the user ID for Twitter handle @united).
2. Get up to the 800 most recent tweets (based on API request caps) from the airline company's Twitter account, where the request response includes organic tweets and retweets and exclude replies[3].

In total, 67,540 tweets were downloaded on December 4, 2022; non-English tweets were filtered to leave 57,404 in total for analysis. Figure 1 shows the distribution of number of tweets (in English language only) per airline account. 74% of airline Twitter accounts returned at least 700 tweets. Generally, the Twitter V2 API allows for downloading up to the 800 most recent tweets (and retweets, which are exclusive of the 800 number) when requesting for tweets without replies, so accounts for which less than exactly 800 tweets were downloaded are accounts that do not have more than 800 organic tweets. As each airline company's Twitter account returned its own tweets, each account has its own minimum and maximum tweet creation date. Overall, the earliest tweet is from March 3, 2011 (which may correspond to an airline company that has not produced more than 800 organic tweets between March 2011 and December 2022), and the latest tweet is from December 5, 2022.

[1] https://github.com/rkhan15/airline-topic-modeling

[2] https://developer.twitter.com/en/docs/twitter-api

[3] Replies were excluded from requests to exclude [automated] customer service actions via tweet replies to customer Twitter accounts, which would not count towards the marketing campaigns/messages airline companies release in the form of organic tweets.

3.2. Data preprocessing

All tweets were converted into a tabular dataset, consisting of the text and the metadata associated with each tweet. Examples of metadata include date and time created, or the number of likes, retweets, or replies, etc. Preprocessing of the tweets was done to clean the raw text by removing text related to links, retweets (i.e. tweet text beginning with “RT”), mentions, hashtags, punctuation, and numbers. Following this, tokenization was performed to split each tweet into individual terms. These terms were then lemmatized to create a set of unigrams, bigrams, and trigrams that make up all of the terms for which the LDA topic model calculates document-term and term-topic probabilities.

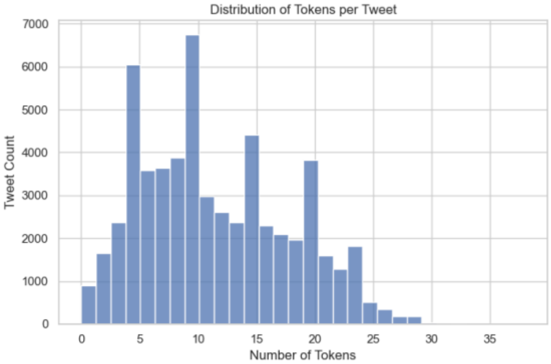


Figure 3: Distribution of number of tokens per tweet, across all 59,404 tweets (in the English language) downloaded for 85 airlines.

Before building the language model, it is useful to understand the distribution of tweet length; this translates into observing the distribution of tweet tokens. Figure 2 shows that number of tokens per tweet has a slight positive skew; based on the interquartile range of the distribution, most documents (tweets) being fed into the LDA model have anywhere from 6 to 16 tokens.

4. Methods

4.1. Topic modeling using Latent Dirichlet Allocation (LDA)

Before creating the LDA model, a dictionary that maps each term in the corpus to an identifier and a corpus that describes the term-tweet frequency are generated.

Perplexity serves as an intrinsic evaluation metric to easily measure the quality of a language model. It is applied on a holdout set of the document dataset and is a function of the probability that a language model assigns to the test corpus [10]. The lower a language model’s perplexity score, the better the model’s [lack of] perplexity.

Topic coherence, or simply coherence, measures how similar words with high probabilities of belonging to a topic are to one another, serving as a measure to define a language model’s semantic interpretability [11]. Among a variety of coherence measures, the UMass coherence [12] measure was selected instead of the default CV coherence measure in gensim’s models.CoherenceModel module (CV is not recommended for coherence measurement[4]); it is based on a smoothed log transform of document-level token cooccurrence counts divided by single token occurrences. The higher an LDA model’s UMass coherence score, the better the model’s coherence.

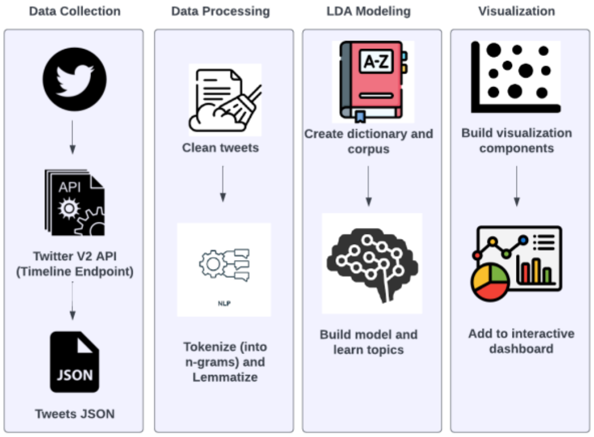
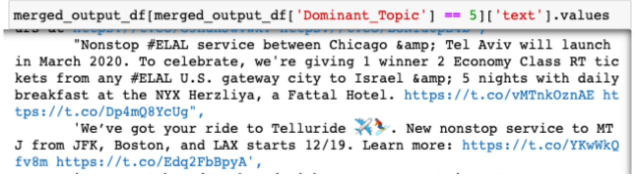

Data Collection
Data Processing
LDA Modeling
Visualization
Twitter V2 API (Timeline Endpoint)
API
JSON
Tweets JSON
Clean tweets
NLP
Tokenize (into n-grams) and Lemmatize
A-Z
Create dictionary and corpus
Build model and learn topics
Build visualization components
Add to interactive dashboard

Figure 2: System design for project.



```
merged_output_df[merged_output_df['Dominant_Topic'] == 5]['text'].values
       "Nonstop #ELAL service between Chicago &amp; Tel Aviv will launch
 in March 2020. To celebrate, we're giving 1 winner 2 Economy Class RT tic
 kets from any #ELAL U.S. gateway city to Israel &amp; 5 nights with daily
 breakfast at the NYX Herzliya, a Fattal Hotel. https://t.co/vMTnkOznAE ht
 tps://t.co/Dp4mQ8YcUg",
       'We’ve got your ride to Telluride ✈️⛷️. New nonstop service to MT
 J from JFK, Boston, and LAX starts 12/19. Learn more: https://t.co/YKwWkQ
 fv8m https://t.co/Edq2FbBpyA',
```

Figure 5: Example of observing tweets with a particular dominant topic; in this case, the dominant topic corresponds to non-stop flight-related services.

5. System Overview

The system design for this project can be broken into the four components, as described in Figure 3.

The data processing component and the LDA modeling technique make use of the gensim[5] package in Python. For tokenization and creation of bi-grams and tri-grams, the models.phrases utility package from gensim is used for

[4] https://github.com/dice-group/Palmetto/issues/13#issuecomment-371553052

[5] radimrehurek.com/gensim.

phrase (collocation) detection. The corpora.Dictionary utility package from gensim is used to create a dictionary object to be fed into the LDA model. The models.CoherenceModel utility package from gensim is used to calculate the coherence metric for different iterations of the LDA topic model.

The visualization component of the project consists of a dashboard using Dash and Plotly components. The intended user of this dashboard in a real-world setting is an analyst who is understanding marketing trends in the airline industry or for a specific airline of choice. The benefits of this dashboard are in its interactivity in understanding how and to what extent different terms (n-grams) contribute to the topic model.

6. Experiments

6.1. Selection of number of topics

In order to determine the best number of topics to ask the model for, a grid-search check, as shown in Figure 4, was conducted to calculate the resulting perplexity and coherence measures of LDA models configured to have a number of topics between 4 and 30. As such, a total of 27 models were trained to assess each based on perplexity and UMass coherence (Figure 4). All models were trained with a random seed for ease of reproducibility. The LDA model with 17 topics was selected to move forward with analysis, as this model strikes a balance between having a) lower perplexity (-20.211587) than other models with less topics, b) a UMass coherence (-8.488740) that is average relative to models with less or more topics, and c) a number of topics that is manageable for a first iteration of reviewing the contents of each topic (to determine a common theme).

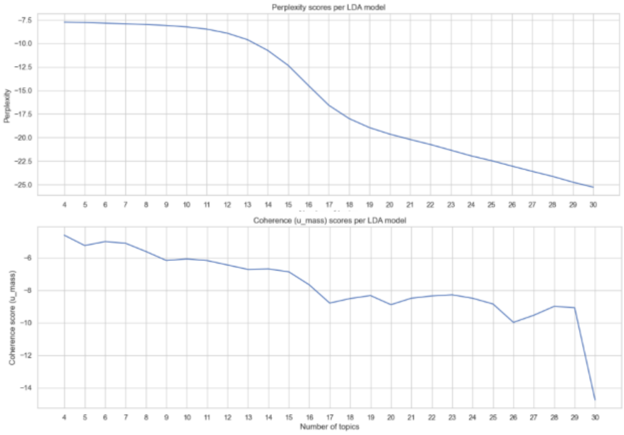


Figure 4: Perplexity and coherence scores for LDA models with varying numbers of topics.

6.2. Selection and naming of topics

After selecting the 17-topic LDA model, each topic was observed by 1) examining each topic's most important or contributing keywords as determined by the weights of the topic and 2) examining tweets for which a particular topic was a tweet's dominant topic. Figure 5 demonstrates an example of observing which tweets were classified by a particular topic according to its dominant topic, which was the topic the LDA model found the highest probability of assigning to the tweet.

Of the 17 topics returned from the model, 13 topics were extracted for further analysis, while the remaining 4 topics were excluded from analysis due to a lack of intra-topic tweet relevance solely based on human interpretability. Following this, the remaining 13 topics were assigned a human-interpretable topic to summarize the tweets for which the topic was a dominant topic. Figure 6 shows the distribution of tweets as labeled by their dominant topic.

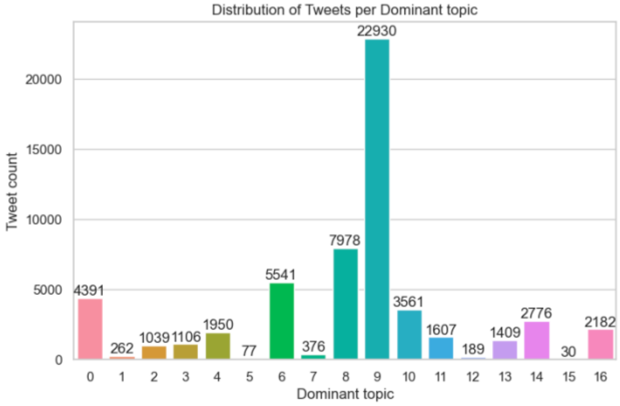


Figure 6: Distribution of tweets as classified by their dominant topic, according to the LDA model.

Additionally, any tweets whose dominant topic was one of the 4 removed topics was also removed from the analysis, bringing the total number of tweets under further analysis to 33,806. The noticeably large reduction in tweets is largely attributed to topic with index 9, under which 22,930 tweets were assigned as having the topic as its dominant topic. As this topic was assigned tweets that did

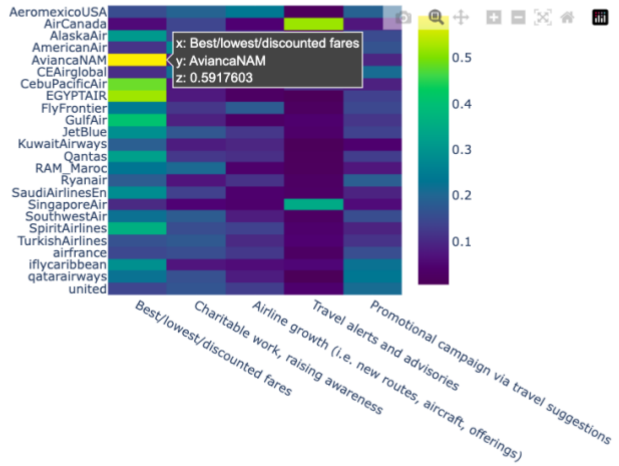


Figure 7: Heatmap showing proportions of topic contributions (bottom axis) to each airline company's set of extracted tweets (left axis). Demonstrated for a sample of the 85 total airlines and a sample of the 13 total topics.

not follow a common, human-interpretable theme, its tweets were removed from further analysis.

Table 1 lists the mapping between topic numeric labels and the human-created and human interpretable names assigned to each of the 13 retained topics.

Topic index	Name
0	Charitable work, raising awareness
2	Future-forward campaigns (i.e. aviation, sustainability)
3	Travel spots with the best views
4	Campaign with a contest
5	Non-stop flight service
6	Promotional campaign via travel suggestions
8	Best/lowest/discounted fares
10	Airline growth (i.e. new routes, aircraft, offerings)
11	Scenic/peaceful travel destinations
12	Flight seating (i.e. group seating, upgrades)
13	Travel alerts and advisories
14	Visit our website for more info
16	Tips for comfortable travel experience

Table 1: Human-interpretable labels assigned to topics from LDA model.

7. Analysis

7.1. Topic contributions to airline tweets

After each tweet was assigned to its dominant topic, it is treated as representative of that topic. Using this, the frequency of tweets per topic from a particular airline were used to compute topic contributions to each airline's set of tweets. As a result, each airline had a percentage contribution from at least 1 of the 13 topics, and these topic-contribution scores are used to determine which topics an airline's tweets correspond to more over other topics. Figure 7 shows how this analysis was visualized in this project's resulting dashboard visualization.

As shown in Figure 7, each airline company can be distinguished by its variety of topic contributions to its tweets. For example, the topic "Best/lowest/discounted fares" summarizes a significant proportion of tweets for several of the airlines shown in the figure. However, some airlines stand out specifically from others for having significant contributions from other topics. United Airlines (@united)'s tweets have a larger share of tweets corresponding to the airline company's growth, while large proportions of Air Canada (@AirCanada)'s tweets and Singapore Airlines (@SingaporeAir)'s tweets comprise travel-related alerts and advisories.

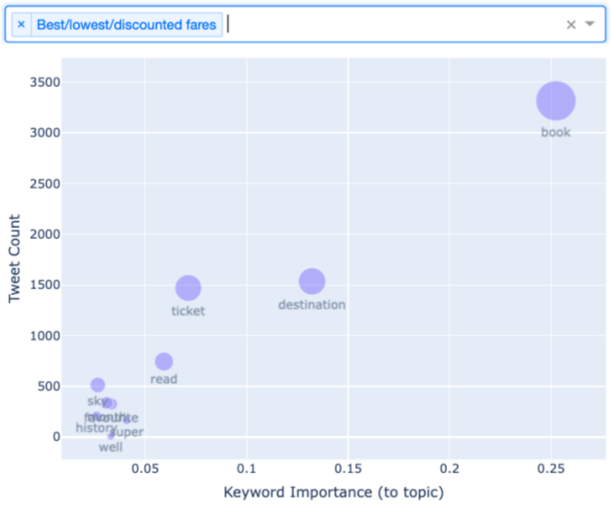


Figure 8a: Topic keywords for topic "Best/lowest/discounted fares" compared by topic tweet count, importance of the keyword to the topic (according to the LDA model), and average engagement with topic tweets containing the keywords (bubble size).

7.2. Topic keyword analysis: Frequency, importance, and average tweet engagement

The LDA topic returns keywords for each topic, which essentially contribute the most to the content classified under the topic. Each topic's keywords were further analyzed to better understand their intra-topic measures such as frequency, importance, and average tweet engagement of tweets containing the keyword.

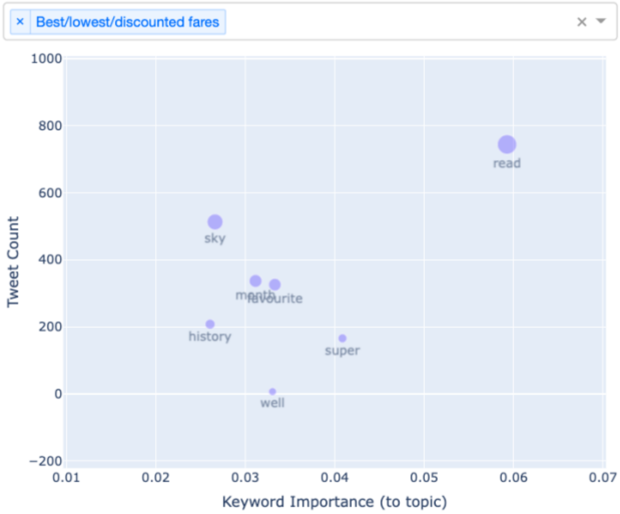


Figure 8b: Zoomed-in crop of keywords shown in the bottom-left corner of Figure 8a.

Frequency was measured by the count of topic tweets that contain each keyword. Keyword importance comes directly from the LDA model's topic descriptions. Average keyword-tweet engagement was derived by first extracting engagement dimensions from the metadata of each tweet, normalizing each measure by the average user engagement under each dimension, summing the normalized engagement measures, and then taking the average over all tweets under that topic-keyword. Here, tweet engagement is a sum of the normalized count of likes, normalized count of retweets, normalized count of replies, and normalized count of quotes. The user-specific normalization was introduced to account for the varying levels of engagement that each airline company may have on the average number of likes, retweets, replies, and quotes per tweet per airline.

Overall, each topic demonstrates a semi-linear relationship between its keywords' importances to the topic and the frequency of tweets containing those keywords. As for all of the topics, the average engagement for topic-keywords also varies especially with keyword tweet counts, which makes sense under the assumption that keywords' average engagement is determined by the frequency of tweets with the keyword, which consequently allow the user to engage more with tweets containing the keyword.

Observing the specific example shown in Figure 8, keywords like "book," "ticket," and "destination" (Figure 8a) contribute the most to the "Best/lowest/discounted fares" topic across all three dimensions of analysis (frequency, importance, and engagement). These keywords intuitively coincide with the keywords one can expect to see beyond Twitter as well that may drive engagement with the tweet and perhaps lead to "booking" the flight, "buying" the ticket, or flying to one's dream "destination" at a low price.

8. Conclusion

Airline companies promote themselves in unique ways on Twitter, and this topic modeling analysis demonstrates that airline marketing strategies cater to the modern traveler. By using Twitter to engage with their customers, they not only promote their unique offerings and enticing prospective travelers with destination offerings and lower fares than competitors, but they also keep the traveler practically engaged through travel alerts as well as morally engaged by promoting their charitable work for organizations in need.

The LDA topic modeling presents two outputs that allow for the interpretability of this analysis. The document-topic matrix allows for the mapping of tweets to their dominant topics, while the topic-term matrix allows for grasping the contribution of specific keywords to each topic based on each keyword's frequency, importance, and average engagement associated with the topic-tweets containing the keyword.

The work in this project could be scoped to focus on a specific time frame to understand seasonality and trend amongst the topics; such work can benefit airline company marketing teams and third-party analysts to justify the creation or adjustment of marketing strategies or to perform competitor-topic analyses across multiple airlines along selected dimensions (i.e. regions supported by the airline).

References

[1] "Stop Calling Your Airline, Message Them on Twitter Instead." https://thriftytraveler.com/guides/airlines/airlines-twitter/ (accessed Dec. 06, 2022).

[2] "Europe's largest airline is a troll on social media — and it's working for them," *Washington Post*. [Online]. Available: https://www.washingtonpost.com/travel/2022/10/04/ryanair-twitter-strategy-gen-z/

[3] B. K. P. D. Balakrishnan, M. I. Dahnil, and W. J. Yi, "The Impact of Social Media Marketing Medium toward Purchase Intention and Brand Loyalty among Generation Y," *Procedia - Social and Behavioral Sciences*, vol. 148, no. 148, pp. 177–185, Aug. 2014, doi: 10.1016/j.sbspro.2014.07.032.

[4] L. Chamberlain, "How Airlines' Embrace Of Social Media Is Evolving After A Decade Of Learning," *Kambr Media*, Feb. 25, 2020. https://www.kambr.com/articles/how-airlines-embrace-of-social-media-is-evolving-after

[5] A. Punel and A. Ermagun, "Using Twitter network to detect market segments in the airline industry," *Journal of Air Transport Management*, vol. 73, pp. 67–76, Oct. 2018, doi: 10.1016/j.jairtraman.2018.08.004.

[6] A. Rane and A. Kumar, "Sentiment Classification System of Twitter Data for US Airline Service Analysis," *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, Jul. 2018, doi: 10.1109/compsac.2018.00114.

[7] D. Blei, B. Edu, A. Ng, M. Jordan, and J. Edu, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003, [Online]. Available: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[8] H. M. Alash and G. A. Al-Sultany, "Improve topic modeling algorithms based on Twitter hashtags," *Journal of Physics: Conference Series*, vol. 1660, p. 012100, Nov. 2020, doi: 10.1088/1742-6596/1660/1/012100.

[9] H.-J. Kwon, H.-J. Ban, J.-K. Jun, and H.-S. Kim, "Topic Modeling and Sentiment Analysis of Online Review for Airlines," *Information*, vol. 12, no. 2, p. 78, Feb. 2021, doi: 10.3390/info12020078.

[10] K. Arora and A. Rangarajan, "Contrastive Entropy: A new evaluation metric for unnormalized language models." Accessed: Dec. 06, 2022. [Online]. Available: https://arxiv.org/pdf/1601.00248.pdf

[11] M. Röder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, 2015, doi: 10.1145/2684822.2685324.

[12] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing Semantic Coherence in Topic Models," ACLWeb, Jul. 01, 2011. https://aclanthology.org/D11-1024/ (accessed Dec. 16, 2022).